Regression-model-final-project

by fabiobianco

04/06/2017

```
library(ggplot2)
library(GGally)
library(datasets)
```

Executive Summary

In this paper we are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome) for the `mtcars` dataset. We are particularly interested in the following two questions:

- Is an automatic or manual transmission better for MPG
- Quantify the MPG difference between automatic and manual transmissions

Exploratory data analysis

```
data("mtcars")
dim(mtcars)
```

```
## [1] 32 11
```

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

```
str(mtcars)
```

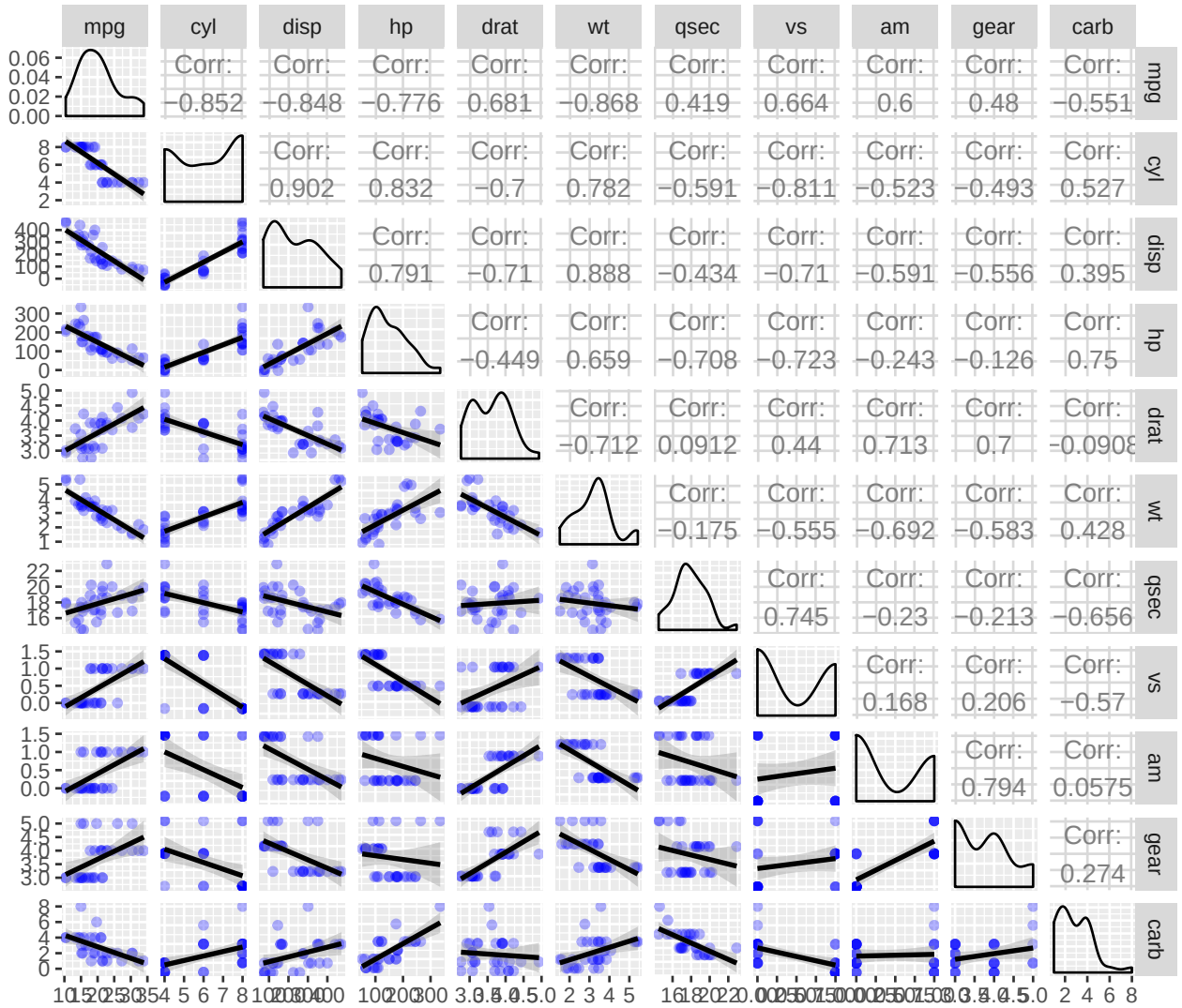
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Now we display the pairwise relations between variables in the `mtcars` dataset

```
g = ggpairs(mtcars, lower = list(continuous = wrap("smooth", method = "lm", alpha = 0.3, color = "blue")))
g
```



In the first columb we can see the relation between the `mpg` variable (outcome) and the other variables (predictors) it seems to be a linear relation (with a positive or negative slope) for every row/predictor and outcome.

Now we explore the relation between the `mpg` variable (outcome) and the `am` variable (predictor)

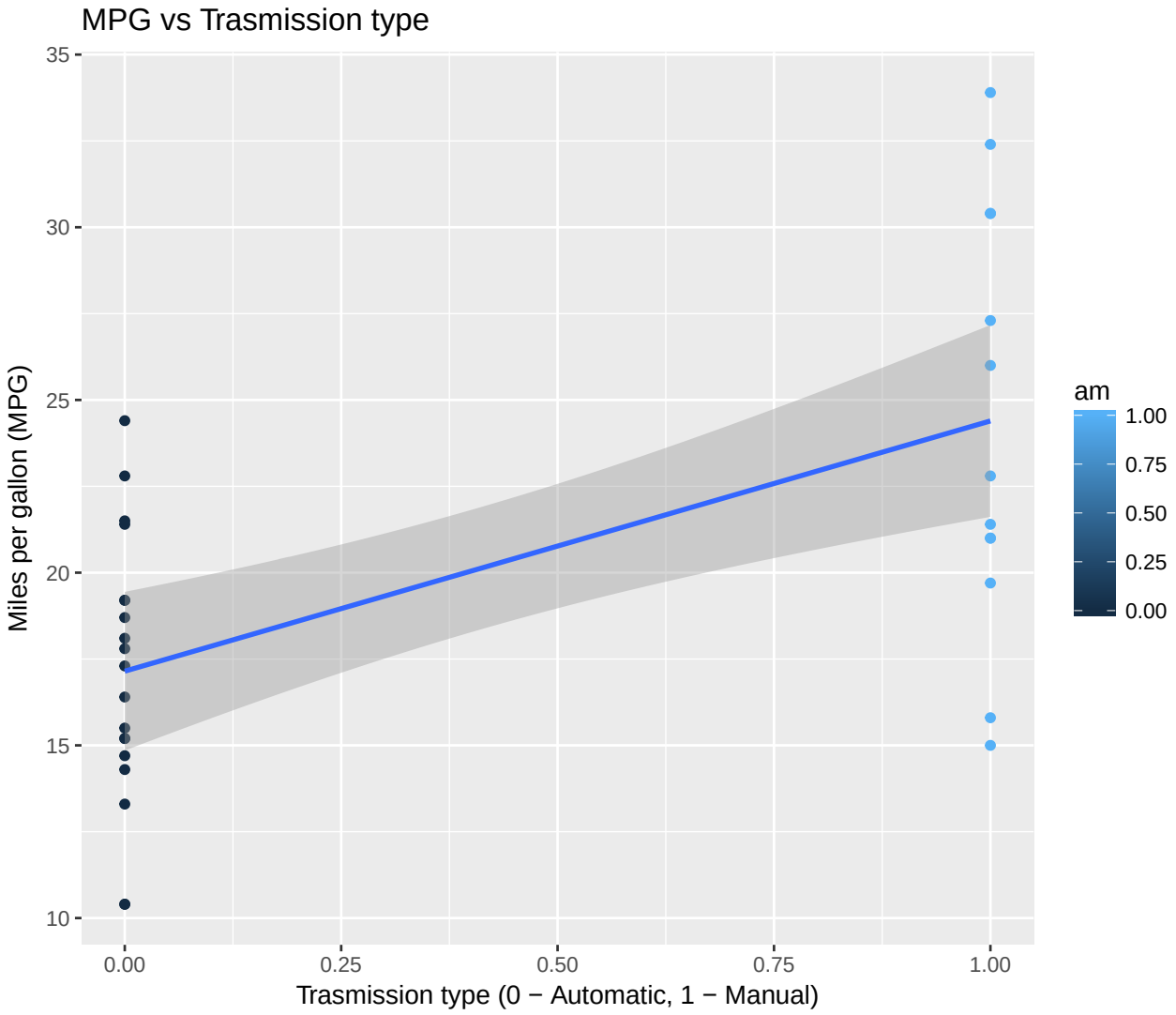
```
mtcars1 <- mtcars[mtcars$am == 1,] # mean MPG for manual trasmission system
summary(mtcars1$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

```
mtcars0 <- mtcars[mtcars$am == 0,] # mean MPG for automatic trasmission system
summary(mtcars0$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
```

```
g = ggplot(mtcars, aes(am, y = mpg, color = am)) + geom_point() + geom_smooth(method = "lm")
g = g + xlab("Trasmission type (0 - Automatic, 1 - Manual)") + ylab("Miles per gallon (MPG)")
g = g + labs(title = paste("MPG vs Trasmission type"))
g
```



Regression Models

Now fit a multivariable linear regression model for the `mtcars` dataset

```
fitall <- lm(mpg ~ . , data = mtcars)
summary(fitall)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

The `fitall` model accounts for 81% of the variance as noted by the adjusted Rsquared value.

Now we investigate the relationship between trasmission type `am`(predictor) and miles per gallon `MPG` (outcome)

```
fit <- lm(mpg ~ am,  data = mtcars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

The `fit` model accounts for 34% of the variance as noted by the adjusted Rsquared value. There seems other predictors have some impact on MPG. The `fit` model predicts an extra 7.245 mpg consuption for manual trasmission veichle versus automatic trasmission veichle. Examining the regression output value, we can see that the p-value for `am` is very clode to zero, indicating there is strong evidence that the coefficient is different fro zero when using this one-variable model.

As a final step we search the model that best fit the data.

```r
summary(bestmodel)
```

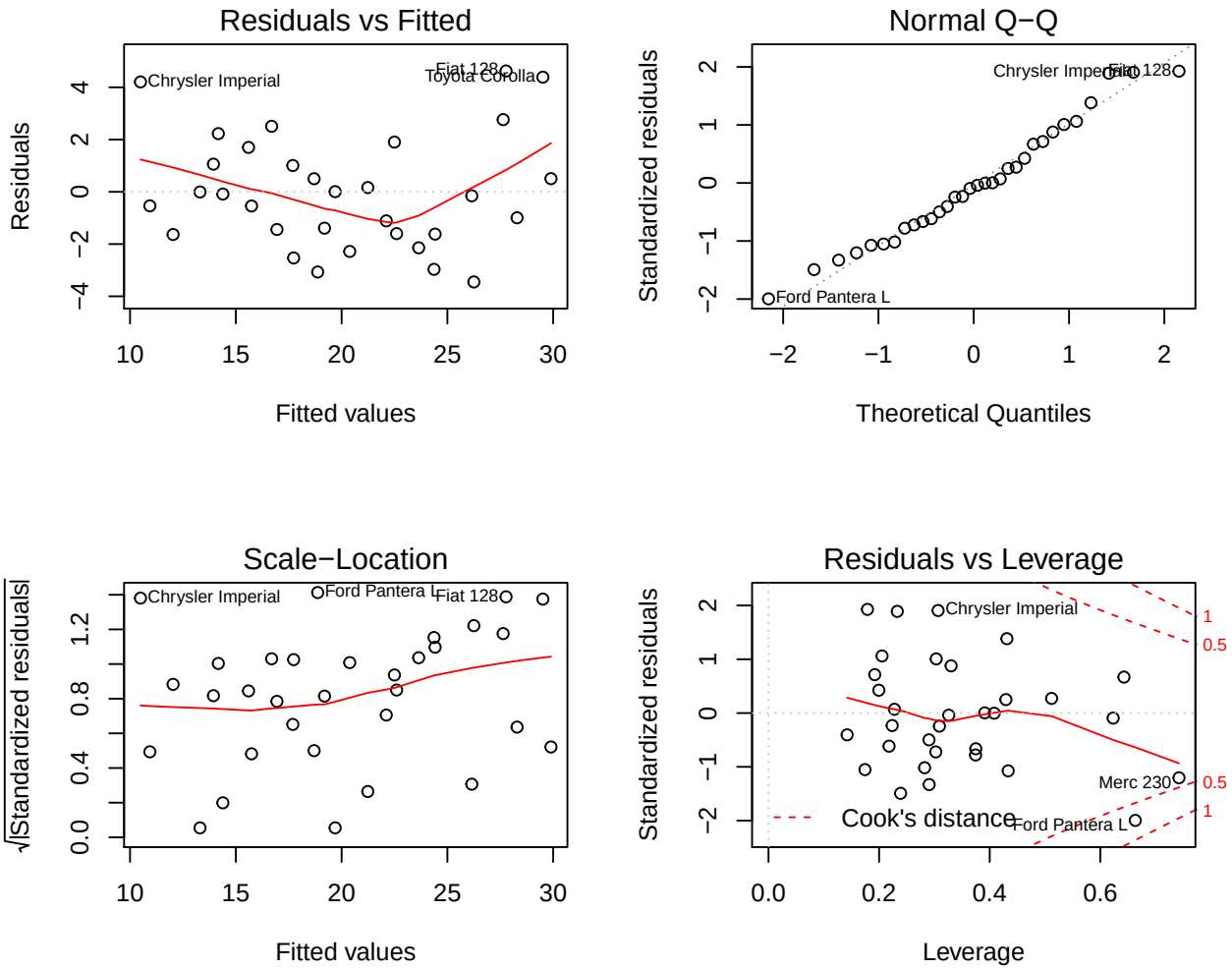
```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

The `bestmodel` accounts for 83% of the variance as noted by the adjusted Rsquared value.

Appendix

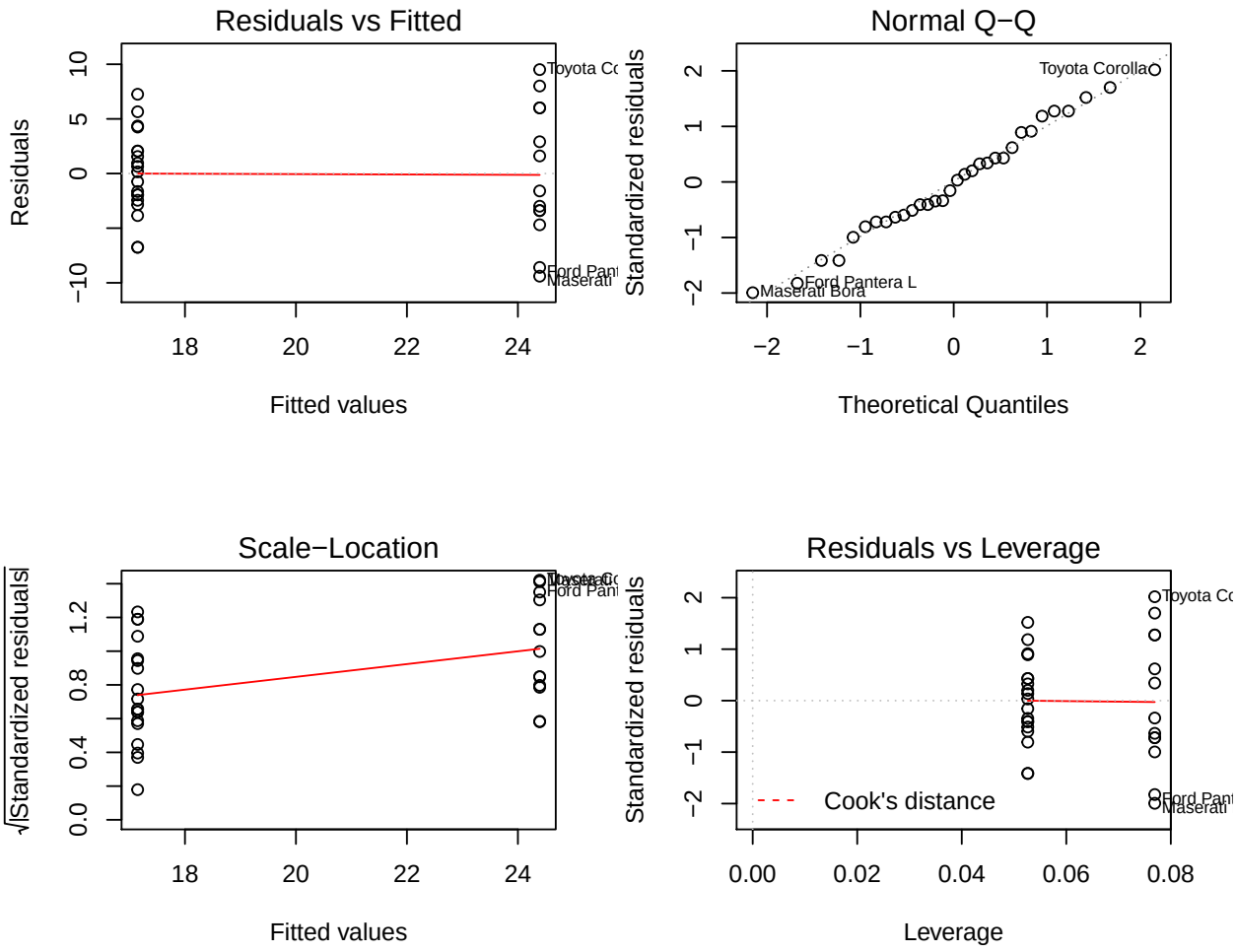
Residual for **Full model** (all predictors)

```r
par(mfrow=c(2, 2))
plot(fitall)
```



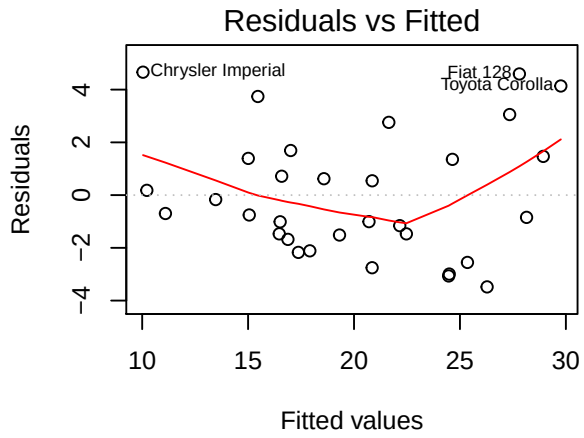
Residual for **single-variable model** (only `am` predictor)

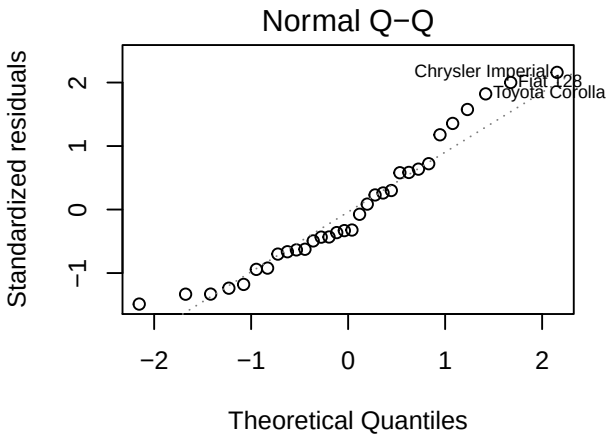
```
par(mfrow=c(2,2))
plot(fit)
```

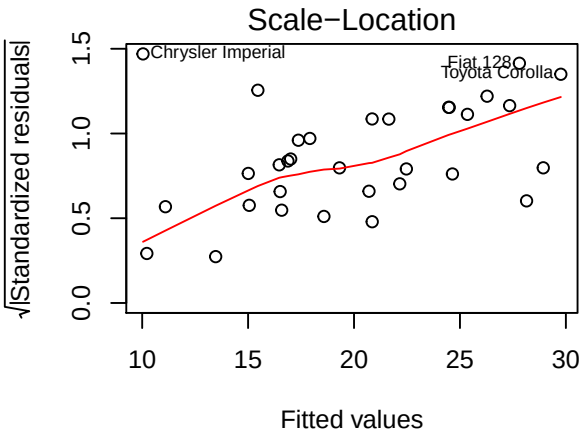


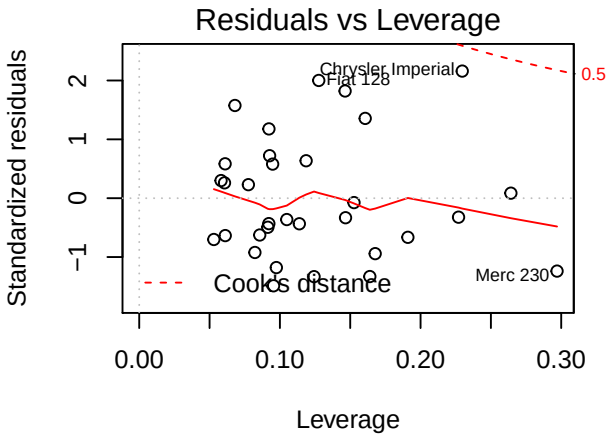
Residual for **best model** (only `am`, `wt`, `qsec` predictors)

```
par(mfrow=c(2, 2))
plot(bestmodel)
```


Residuals vs Fitted
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals
Fitted values


Normal Q−Q
Chrysler Imperial
Fiat 128
Toyota Corolla
Standardized residuals
Theoretical Quantiles


Scale−Location
Chrysler Imperial
Fiat 128
Toyota Corolla
√|Standardized residuals|
Fitted values


Residuals vs Leverage
Chrysler Imperial
Fiat 128
Merc 230
0.5
Cook's distance
Standardized residuals
Leverage